EPOC: a survival perspective Early Pattern detection model for Outbreak Cascades

Chaoqi Yang, Qitian Wu, Xiaofeng Gao*(✉), and Guihai Chen

Shanghai Key Laboratory of Scalable Computing and Systems,
Department of Computer Science and Engineering, Shanghai Jiao Tong University,
Shanghai, 200240, P.R.China
ycqsjtu@gmail.com, echo740@sjtu.edu.cn, {gao-xf,gchen}@cs.sjtu.edu.cn

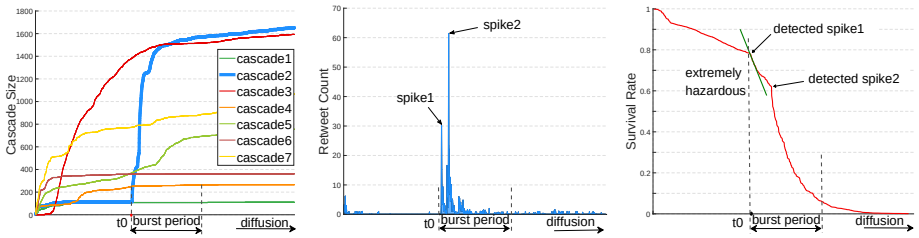
Abstract. The past few decades have witnessed the booming of social networks, which leads to a lot of researches exploring information dissemination. However, owing to the insufficient information exposed before the outbreak of the cascade, many previous works fail to fully catch its characteristics, and thus usually model the burst process in a rough manner. In this paper, we employ survival theory and design a novel survival perspective Early Pattern detection model for Outbreak Cascades (in abbreviation, EPOC), which utilizes information both from the static nature and its later diffusion process. To classify the cascades, we employ two Gaussian distributions to get the optimal boundary and also provide rigorous proof to testify its rationality. Then by utilizing both the survival boundary and hazard ceiling, we can precisely detect early pattern of outbreak cascades at very early stage. Experiment results demonstrate that under three practical and special metrics, our model outperforms the state-of-the-art baselines in this early-stage task.

Keywords: Early-stage Detection · Outbreak Cascade · Survival Theory · Cox's Model · Social Networks

1 Introduction

The rapid development of modern technology has changed the lifestyles to a large extent compared to a few years ago. Every day millions of people express ideas and interact with friends through online platforms like Twitter and Weibo. On these platforms, registered users are able to *tweet* short messages (e.g., up to 140 characters in Twitter), and others who are interested in it will give likes, comments, or more commonly, *retweets*. Such retweeting would potentially disseminate and further spread information to a large number of users, which forms a *cascade* [1]. While the cascade grows larger and get more individuals involved, a sudden *burst* will definitely arrive, which we call a *spike*. As a matter of fact, detecting and predicting the burst pattern of a cascade, especially at early stage, attract lots of attention in various domains: meme tracking [2], stock bubble diagnosis [3], and sales prediction [4], etc.

* Corresponding author.



(a) Cascade Life Cycle (b) Retweeting @Cascade2 (c) Survival Curve @Cascade2

Fig. 1. Samples of Cascade Diffusion on Twitter

However, to fully understand the burst pattern of cascades ahead of time will meet three major challenges. **First** and foremost, due to the deficiency of available information and its disorder nature at early stage [5], one can hardly catch distinguishing signs on whether a cascade will break out. The **second** challenge stems from the significantly distinct life span of different cascades [6], which makes it tough to extract typical features. Worse still, this distinctiveness makes it hard for researchers to set suitable observation time, owing to the variety of life spans. The **third** challenge is that the burst pattern of cascades usually follows a quick *rise and fall* law [7], which lasts a few minutes but causes magnificent influence. In this situation, the correlations between the history and the near future can be hardly characterized by traditional models.

Shown in Figure 1(a), we plot the diffusion process of seven real-world cascades from Twitter. We can see that @Cascade2 shares almost the same pattern with @Cascade1 before it outbreaks at time t_0, which means that it is hard for us to catch the distingushing signs using the early information. As the second challenge states, @Cascade1∼7 represent different life span at early stage. While @Cascade6 ends its diffusion, @Cascade3 is just about to start propagation, and it still enlarges even at the end of observation. The third challenge can be vividly described in Figure 1(b), where we focus on @Cascade2 and plot how it is retweeted. Figure 1(b) shows that @Cascade2 experiences a mild propagation when it appears, but after time t_0, it goes through two large retweeting spikes (sudden falls in survival curve ploted in Figure 1(c)), and the final amount of retweeting explodes to about 1600 during the burst period. These three core challenges motivate us to design a model that can handle this quick rise and fall pattern, characterize different cascades uniformly, and detect the burst pattern as early as possible.

Motivated by the study of death in biological organisms, in this paper, we regard the diffusion of cascades as the growing process of biological organisms. Since Cox's model is widely used to characterize the life span of biological organisms, here we adopt Cox's model with the knowledge of cascades, transforming the burst detection task into diagnosis of cascade life table, and then we build a survival perspective Early Pattern detection model for Outbreak Cascades, in abbreviation, EPOC. Though previous work [8] has also tried Cox's model, their

work is mainly based on unsubstantiated observations as well as only taking one feature into consideration, which does not address the above challenges at all.

In our EPOC, to consider the influential factors from different perspectives, we harness three features from each cascade (retweet sequence, follower number sequence, and original timestamp) to capture the effectiveness of temporal information [9], the influence of involved users [10], and the dynamics of user activity [11]. Then, to study the distinctiveness of cascades' life span, we train an effective Cox's model and employ two Gaussian distributions to fit the survival probability of viral and non-viral cascades at different time point respectively, and obtaining a survival boundary between the viral and the non-viral, which is further proven to be well-defined theoretically. Finally, as the static and dynamic nature of cascade diffusion are both important indicators of cascade virality, we jointly consider survival probability and hazard rate, which considerably enhances our model's performance in handling the quick rise and fall pattern. We then employ three special metrics (K-coverage, Cost, Time ahead) to compare EPOC with two basic machine learning methods (LR, SVR) and three powerful baselines published in recent literatures (PreWhether [12], SEISMIC [10], SansNet [8]) on two large real-world datasets: Twitter and Weibo. Experiment results show that EPOC outperforms these five methods in burst pattern detection at very early stage.

Our main contributions are summarized as:

- We adopt survival theory and establish a powerful burst detection model EPOC for cascade diffusion, which can handle the quick rise-and-fall pattern as well as the significantly distinct life span of cascades at the early stage.
- We utilize both static and dynamic information from cascades, obtain a dimidiate boundary with two Gaussian distributions, and then novelly use the burst pattern to help predict the popularity of an online content.
- We adopt three special metrics and conduct extensive experiments on two large real-world data sets (Twitter and Weibo). The results show that EPOC gives the best performance comparing with five state-of-the-art approaches.

The remainder of the paper is organized as follows. Some common notions of survival theory and the basic Cox's model are introduced in Section 2. The design of our proposed model EPOC is specified in Section 3. We evaluate and analyze our model on Twitter and Weibo in Section 4. We review several related works in Section 5. Finally, we conclude our work and highlight the possible future perspectives in Section 6.

2 Survival Analysis and Cox's Model

In this section, we give some definitions about survival theory in social networks. Initially, when a user shares the content with her set of friends, several of these friends share it with their respective sets of friends, and a *cascade* of resharing can develop [13]. Once the size of this cascade grows above a certain *threshold* ρ, we call it goes *viral*, and otherwise *non-viral*. To quantitively describe these

statues of cascade diffusion, we introduce *survival function* and *hazard function* respectively in Definition 1 and Definition 2.

Definition 1. *(Survival Function): let $S(t) \in (0,1)$ denote the survival probability of cascade subject to time t, i.e., at time t, cascade has the probability of S(t) to be non-viral, where $S(t)$ is naturally monotonic decreasing with time t.*

Definition 2. *(Hazard Function): let $h(t) \in (0,\infty)$ denote the hazard rate of cascade at time t on the condition that it survives until time t ,i.e., h(t) is the negative derivative of survival probability $-\frac{\mathrm{d}S(t)}{\mathrm{d}t}$ to the survival function $S(t)$, specifically given by the following formula,*

$$h(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} \cdot \frac{1}{S(t)}. \tag{1}$$

Since Cox's survival model was proposed [14], it has been widespread used in the analysis of time-to-event data with censoring and covariates [15]. In this work, we use Cox's proportional hazard model with time-dependent covariates (also called Cox-extended model) to characterize the association between early information and the cascade statues (viral or non-viral).

Basic Model: for cascades $i = 1, 2, \cdots, n$, they share the same baseline hazard function denoted as $h_0(t)$, and $\boldsymbol{X_i}(t) = \{x_1^{(i)}, x_2^{(i)}, \cdots, x_m^{(i)}\}$ denotes the feature vector of the i_{th} cascade, where $h_0(t)$ does not depend on each $\boldsymbol{X_i}(t)$ but only on t. $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_m\}$ is the parameter vector of our hazard model. We specify the hazard function of i_{th} cascade as follows,

$$h_i(t) = h_0(t) \cdot \exp\left(\boldsymbol{\beta}^T \boldsymbol{X_i}(t)\right). \tag{2}$$

Because the model is proportional, i.e., given i_{th} and j_{th} cascade, the relative hazard rate $\lambda_{i,j}$ can be concretely given by,

$$\lambda_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \cdot \exp\left(\boldsymbol{\beta}^T \boldsymbol{X_i}(t)\right)}{h_0(t) \cdot \exp\left(\boldsymbol{\beta}^T \boldsymbol{X_j}(t)\right)} = \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{X_i}(t)\right)}{\exp\left(\boldsymbol{\beta}^T \boldsymbol{X_j}(t)\right)} \tag{3}$$

where $\boldsymbol{\beta}$ is the parameter vector, $\boldsymbol{X_i}(t)$ and $\boldsymbol{X_j}(t)$ are respectively the feature vectors of i_{th} and j_{th} cascade. From Eqn. (3), it is easy to conclude that the baseline hazard does not play any role in relative hazard rate $\lambda_{i,j}$, i.e., the model is also a semi-parametric approach. Therefore, instead of considering the absolute hazard function, we only care about the relative hazard rate of cascades, which only concerns parameter vector $\boldsymbol{\beta}$. Then we use Maximum Likelihood Estimation to get parameter vector $\boldsymbol{\beta}$. We denote i_{th} cascade time-to-event as t_i, and assume that $0 < t_1 < t_2 < \cdots < t_n$. The Cox's partial likelihood is given by,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{h_i(t_i)}{\sum_{j=i}^{n} h_j(t_i)} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{exp\left(\boldsymbol{\beta}^T X_i(t_i)\right)}{\sum_{j=i}^{n} exp\left(\boldsymbol{\beta}^T X_j(t_i)\right)} \right)^{\delta_i}, \tag{4}$$

where δ_i means whether the data from i_{th} cascade is censored, i.e., if the event happens to i_{th} cascade, then δ_i equals to 1, and otherwise 0. Then the log-partial likelihood of parameter vector $\boldsymbol{\beta}$ can be calculated as,

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}^T X_i(t_i) - \log \left(\sum_{j=i}^{n} \exp \left(\boldsymbol{\beta}^T X_j(t_i) \right) \right) \right], \tag{5}$$

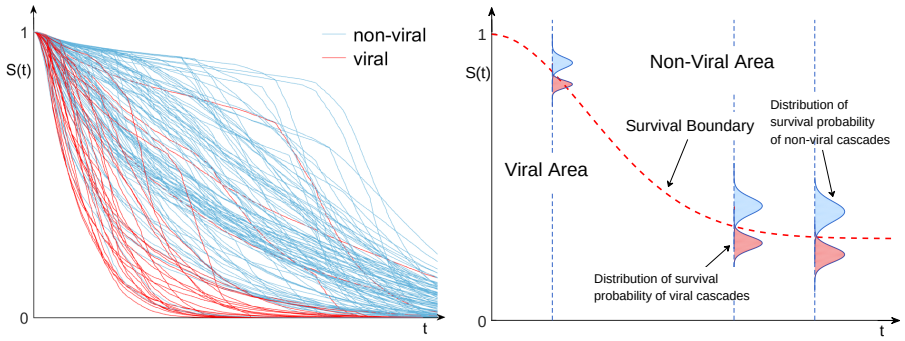
maximizing the log-partial likelihood by solving equation $\frac{\mathrm{d} \log L(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = 0$, then we can get the numerical estimation of parameter vector $\boldsymbol{\beta}$ using Newton method.

3 EPOC: detecting Early Pattern of Outbreak Cascades

Based on the basic model stated previously, in this section, we combine the Cox's model with our knowledge of cascades, and make it suitable to handle the task of detecting the early pattern of outbreak cascades. Here we regard cascades as complex dynamic objects that pass through successive stages as they grow. During this process of growth, the survival probability and the hazard rate of cascades will change dynamically. The high survival probability and low hazard rate suggest that cascades are unlikely to be viral in the future, while the low survival probability as well as high hazard rate imply the opposite. In this sense, we introduce the *survival boundary* and the *hazard ceiling* to help accomplish this challenging task at very early stage.

Feature Selection: as is stated previously, the effectiveness of temporal information, the influence of involved users, and the dynamics of user activity are all powerful indicators of the cascade statues. Therefore, in this experiment, we utilize three features accordingly: *timestamp of each retweet*, *number of followers* of every user involved in the cascade, and *timestamp of the first tweet*.

3.1 Survival Boundary: a Static Perspective



(a) Survival Functions of Cascades (b) Survival Boundary

Fig. 2. Survival Functions and Survival Boundary

To detect the early pattern of outbreak cascades, firstly, we characterize the survival functions of all cascades. Shown in Figure 2(a), the red lines represent the survival functions of viral cascades, and the blue lines show the non-virals'. Then we are supposed to divide the estimated survival functions of all cascades into two classes (viral and non-viral). In other word, we need to find a *survival boundary*. As is illustrated in Figure 2(b), the red dashed line separates the two categories of blue (non-viral cascades) and red (viral cascades).

Previous works [16] have demonstrated that at a fixed observing time t, the distribution of survival probability of different cascades obeys Gaussian distribution. Based on this knowledge, we employ two random variables: f_v^t (for viral cascades) and f_n^t (for non-viral cascades) subject to time t, which satisfy the Gaussian. Formally, we specify this assumption in Definition 3.

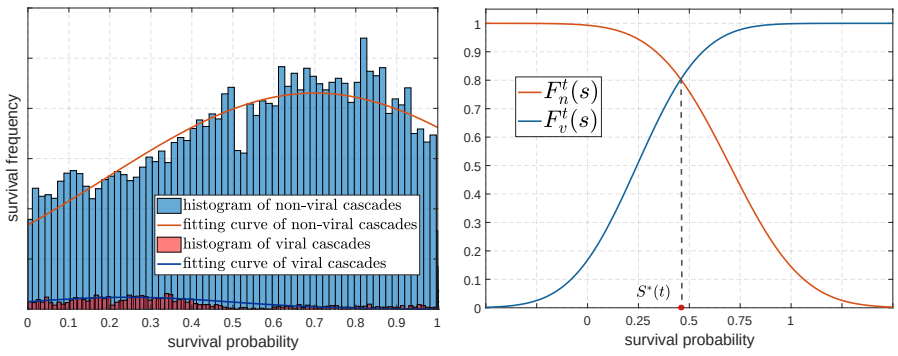
Definition 3. *For any Given time t, we have $f_v^t \sim \mathcal{N}(\mu_v^t, \sigma_v^t)$ and $f_n^t \sim \mathcal{N}(\mu_n^t, \sigma_n^t)$, where μ_v^t, σ_v^t and μ_n^t, σ_n^t are the parameters of Gaussian distribution for viral and non-viral cascades subject to time t.*

Based on Definition 3, for a given time t, the survival probability of viral and non-viral cascades can be respectively characterized as f_v^t and f_n^t. Therefore, the task to find the optimal survival boundary is to give the suitable separation between two Gaussian distributions.

Definition 4. *(Survival Boundary): for any given time t, assume the survival boundary to be $S^*(t)$, which is given by the following formula,*

$$\int_{-\infty}^{S^*(t)} \frac{1}{\sqrt{2\pi}\sigma_v^t} \exp\left(-\frac{(x-\mu_v^t)^2}{2{\sigma_v^t}^2}\right) \mathrm{d}x = \int_{S^*(t)}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_n^t} \exp\left(-\frac{(x-\mu_n^t)^2}{2{\sigma_n^t}^2}\right) \mathrm{d}x. \tag{6}$$

Then the optimal survival boundary can be calculated as $S^*(t) = \frac{\mu_v^t \sigma_n^t + \mu_n^t \sigma_v^t}{\sigma_v^t + \sigma_n^t}$.



(a) Survival Frequency and Fitting (b) Survival Boundary $S^*(t)$

Fig. 3. Survival Frequency and Survival Boundary at Time t

As is shown in Figure 3(a), given time t, we plot the frequency histograms of survival probabilities of both viral and non-viral cascades (blue bars represent non-viral ones, and red bars represent viral ones). Then we use two Gaussian distribution curves f_v^t and f_n^t to fit these two histograms. Next, to simplify our problem, we employ the cumulative distribution function of f_v^t and f_n^t, respectively denoted as $F_v^t(s)$ and $F_n^t(s)$, specifically we have,

$$F_v^t(s) = P(S < s) = \int_{-\infty}^{s} \frac{1}{\sqrt{2\pi}\sigma_v^t} \exp\left(-\frac{(x-\mu_v^t)^2}{2{\sigma_v^t}^2}\right) \mathrm{d}x, \tag{7a}$$

$$F_n^t(s) = P(S > s) = \int_{s}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_n^t} \exp\left(-\frac{(x-\mu_n^t)^2}{2{\sigma_n^t}^2}\right) \mathrm{d}x. \tag{7b}$$

Finally, we plot $F_v^t(s)$ and $F_n^t(s)$ in Figure 3(b), and the x-coordinate of the only intersection $S^*(t)$ is the optimal survival boundary subject to time t.

3.2 Well-Definedness of Survival Boundary

In order to make the problem more complete and rigorous, in this subsection, we mainly discuss the monotonicity of the survival boundary, which is given in Definition 4, i.e., we will prove that the optimal survival boundary is itself a survival function.

In fact, during the observation period, we conclude three solid facts. First of all, the survival probabilities of both viral and non-viral cascades are naturally monotonic decreasing with time t, so the average survival probabilities of both cascades are also monotonic decreasing. Besides, non-viral cascades intuitively possess a higher survival probability, thus the average survival probability for non-viral cascades μ_n^t is reasonably larger than that of viral ones μ_v^t. Further more, real-word data shows that the survival probability range of non-viral cascades appears to be more dynamic and uncertain, which means its relative fluctuation of standard deviation σ_n^t is also larger than σ_v^t. Formally, we specify these three conclusions in Lemma 1.

Lemma 1. *For any given time t, μ_v^t, σ_v^t and μ_n^t, σ_n^t respectively represent the average survival probability and its standard deviation of viral and non-viral cascades. Given time $t' > t$, we have*

$$\begin{cases} \mu_v^t \geq \mu_v^{t'} \\ \mu_n^t \geq \mu_n^{t'} \end{cases}, \quad \mu_n^t \geq \mu_v^t, \quad \frac{\sigma_n^{t'} - \sigma_n^t}{\sigma_n^t} \geq \frac{\sigma_v^{t'} - \sigma_v^t}{\sigma_v^t}, \qquad \forall \;\; 0 < t < t'. \tag{8}$$

Based on Definition 4 and Lemma 1, we given detailed proof that the optimal survival boundary is itself a survival function.

Theorem 1. *The optimal survival boundary $S^*(t)$ is monotonic decreasing with time t, i.e., $S^*(t)$ is also a survival function. Formally, we have*

$$S^*(t) \geq S^*(t'), \qquad \forall \;\; 0 < t < t', \tag{9}$$

Proof. For $\forall\ 0 < t < t'$, we have

$$\begin{aligned} S^*(t) - S^*(t') &= \frac{\mu_n^t \sigma_v^t + \mu_v^t \sigma_n^t}{\sigma_n^t + \sigma_v^t} - \frac{\mu_n^{t'} \sigma_v^{t'} + \mu_v^{t'} \sigma_n^{t'}}{\sigma_n^{t'} + \sigma_v^{t'}} \\ &= \frac{(\mu_n^t - \mu_v^{t'})\sigma_v^t \sigma_n^{t'} + (\mu_v^t - \mu_n^{t'})\sigma_n^t \sigma_v^{t'} + (\mu_v^t - \mu_v^{t'})\sigma_n^t \sigma_n^{t'} + (\mu_n^t - \mu_n^{t'})\sigma_v^t \sigma_v^{t'}}{(\sigma_n^t + \sigma_v^t)(\sigma_n^{t'} + \sigma_v^{t'})} \\ &\geq \frac{(\mu_v^t - \mu_v^{t'})\sigma_v^t \sigma_n^{t'} + (\mu_n^t - \mu_n^{t'})\sigma_v^t \sigma_n^{t'} + (\mu_v^t - \mu_v^{t'})\sigma_n^t \sigma_n^{t'} + (\mu_n^t - \mu_n^{t'})\sigma_v^t \sigma_v^{t'}}{(\sigma_n^t + \sigma_v^t)(\sigma_n^{t'} + \sigma_v^{t'})} \\ &\geq 0, \end{aligned} \tag{10}$$

according to Lemma 1. We can easily conclude that $S^*(t) \geq S^*(t')$.

3.3 Hazard Ceiling: a Dynamic Perspective

As is defined in Definition 2, hazard function is specifically denoted as $h(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} \cdot \frac{1}{S(t)}$, we can easily monitor the hazard function $h(t)$ of a cascade when given its survival function $S(t)$.

To detect the early pattern of outbreak cascades, many previous works usually ignore the underlying arrival process of retweets, instead, they only consider the relationship between the static size of cascade and a predefined threshold[6] [17], then determine whether the cascade is suffering a burst period. However, before the static size of a cascade accumulates to a certain threshold, its burst pattern can be exactly uncovered from dynamic information, such as the hazard function $h(t)$ in this problem. Intuitively, we conclude that if at a certain time t_0, the hazard function $h(t)$ of a cascade suddenly rises above a *hazard ceiling* α, in other word, $h(t_0) > \alpha$, we deem that the burst period of this cascade begins.

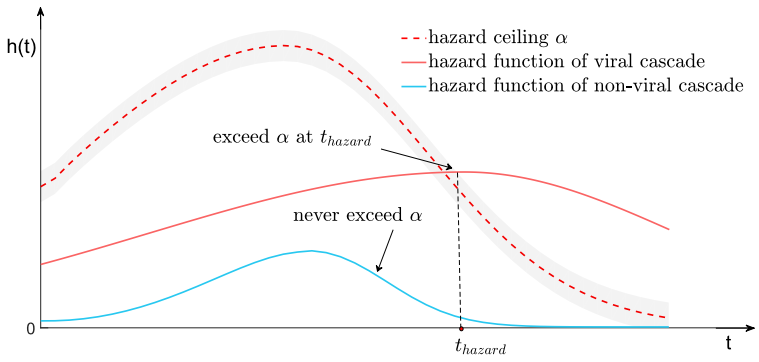


Fig. 4. Hazard Functions and Hazard Ceiling

However, instead of utilizing a fix threshold, we employ the baseline hazard function with a 5% hazard-tolerant interval as hazard ceiling (illustrated in Figure 4), since intuitively the characteristics of cascades may vary a lot during the

diffusion process. In Figure 4, the hazard ceiling is drawn in red dash line with a grey hazard-tolerant interval, and the red solid line and blue solid line respectively denote the hazard functions of a viral cascade and a non-viral cascade. We can clearly conclude that the blue line never exceeds hazard ceiling α, and the red line exceeds α and its hazard-tolerant interval at t_{hazard}. Therefore, we deem that at t_{hazard}, this cascade goes viral and starts to burst.

3.4 Incorporation of Two Techniques

In this subsection, we conclude our method and integrate survival boundary and hazard ceiling. The whole process of EPOC is shown in Alg. 1.

Algorithm 1: Algorithm of EPOC

Input: training data D, test data D', threshold ρ, hazard ceiling α.
Output: status vector V, detect time T.

```
1  Set labels for each cascade from D using threshold ρ ;
2  Train a Cox's model C with time-dependent data D ;
3  Initialize survival function set as S ;
4  foreach d in D do
5  |  estimate the survival function S_d(t) of d using C ;
6  |  add S_d(t) to S;
7  Train an optimal survival boundary S* with S ;
8  foreach d' in D' do
9  |  estimate the survival function S_d'(t) and hazard function h_d'(t) of d' ;
10 |  if S_d'(t) firstly falls down below S*(t) at time t_0 then
11 |  |  add 1 to S ;
12 |  |  if h_d'(t) firstly rises up above α at time t_1 then
13 |  |  |  add min{t_0, t_1} to T ;
14 |  |  else
15 |  |  |  add t_0 to T ;
16 |  else
17 |  |  add 0 to S ;
18 |  |  add none to T ;
19 return S and T.
```

In Alg. 1, *Line1*∼*Line3* is the initialization, and especially we train the Cox's model with time-dependent features in *Line2*. Then the optimal survival boundary is estimated in *Line4*∼*Line7*, after that, we detect the burst pattern between *Line8* and *Line18* using both survival probability and hazard rate.

4 Experiments

In this section, we conduct comprehensive experiments to verify our model in early pattern detection of outbreak cascades. Firstly, we describe the data sets (Twitter and Weibo) and five comparative state-of-the-art baselines in detail. Then we conduct our experiments as well as providing corresponding analysis.

4.1 Data Sets

We implement our model EPOC on two large real-world data sets: *Twitter* and *Weibo*. Twitter is one of the most famous social platforms in the world with annually 0.5 billion users. We densely crawl the tweets that contains hashtags with Twitter search API. In our experiments, a cascade is considered to consist of all tweets with the same hashtag. Another large dataset Weibo is from an online resource[1]. However, different from Twitter, due to the sparsity of hashtags in Weibo, a cascade is defined by the diffusion of a single microblog. More detailed information of two data sets can be found in Table 1.

Table 1. Data sets information

Data set	# of cascades	Type	Range	Year	Size(GB)
Twitter	166,076	hashtag	Aug.13th - Sep.10th	2017	3.827
Weibo	300,000	microblog	Sept.28th - Oct.29th	2012	1.426

4.2 Experiment Setting

For our model implementation, we need to specify some settings. Because large cascades are rare [13], in this paper, we set threshold for viral and non-viral cascades to be 95 percentile in both Twitter and Weibo, where a larger size will be regarded as viral cascade, and otherwise non-viral. As cascades are formed by large resharing activities and can potentially reach a large number of people [13], we only consider the cascades with a tweet count larger than 50 in Twitter and filter out the remains. As for Weibo, the out line is set to be 80.

In the outset of our experiments, we randomly divide each data set into two parts, 80% of the cascades is employed as training data, and the remaining one-fifth as test data. As for the hazard ceiling, in this paper, we use the baseline hazard function as ceiling and set 5% as the hazard-tolerant interval.

4.3 Baselines

From previous literatures, we select a variety of approaches from different perspectives to compare our EPOC: traditional machine learning methods, Bayesian methods, survival methods, and time series methods.

- *Linear Regression (LR)*: Linear regression is a simple and feasible way to characterize the relationship between variables and final result. In this paper, we divide the observation time into twelve time periods, then implement LR with L1 regularization based on different time periods, utilizing the observed information to predict whether or when a cascade goes viral.
- *Support Vector Regression (SVR)*: As is widely used in various areas, SVR is a powerful regression model. We use SVR with Gaussian kernel as a baseline to predict whether a cascade will go viral or even burst in the near future. More detailed implementation of SVR is similar to linear regression.

[1] arnetminer.org/Influencelocality

- *PreWhether* [12]: From a Beyassian perspective, PreWhether is one of the pioneers in social content prediction, which utilizes three temporal features (sum, velocity, and acceleration) to infer the content ultimate popularity. In our experiments, we also use the same time period manner to implement PreWhether.
- *SEISMIC* [10]: SEISMIC is a point process based time series model, which takes individual's influence into consideration. Since the model itself is designed to predict the popularity of single tweets in social networks, we extend it to suit our goals of cascades' burst pattern detection.
- *SansNet* [8]: SansNet is a network-agnostic approach proposed in recent literature, which also regards the burst detection task as a judgement of viral and non-viral. This method shows its detection performance using only the time series information of a cascade.

4.4 Burst Pattern Detection

Burst or Not: to detect the early pattern of outbreak cascades, we primarily divide this problem into two steps. Firstly, we detect whether a cascade will outbreak based on the observed information. Since large cascades are arguably more striking [13], in this classification task, we employ two special metrics: k-coverage and Cost. k-coverage mainly focuses on those cascades with a very large size. Specifically, it is calculated by $\frac{n}{k}$, $(k \geq n)$, where k is the number of the largest cascades being concentrated on, and n denotes the number of cascades we successfully detect from the top-k viral cascades. Here in this work, n equals 50. Cost (more precisely called sensitive cost) is a targeted metric, which is selected to handle the problem of unequal-cost. If a viral cascade (like a rumor [1]) is classified to be non-viral, it will cost a lot when this cascade gets larger and causes a big trouble. On the contrary, if we misclassify a non-viral cascade, it only costs some additional labor. Cost is specified in Eqn. (11),

$$Cost = \frac{FNR \times p \times Cost_{FN} + FPR \times (1-p) \times Cost_{FP}}{p \times Cost_{FN} + (1-p) \times Cost_{FP}}, \tag{11}$$

where FNR is the false negative rate, FPR is the false positive rate, p is the proportion of viral cascades in all cascades, $Cost_{FN}$ and $Cost_{FP}$ are entries in cost matrix. We also specify the cost matrix in Table 2.

Table 2. Unequal-Cost Matrix

Real Class	**Detected Class**	
	Viral	**Non-viral**
Viral	$Cost_{TP} = 0$	$Cost_{FN} = 5$
Non-viral	$Cost_{FP} = 1$	$Cost_{TN} = 0$

Performance Analysis. The results of burst detection are aggregated in Table 3 and the underlined numbers show the best results. One can see that in general, our EPOC performs relatively better than five baselines in terms of both k-coverage and Cost. LR also shows great performace in k-coverage on Weibo,

and it works much better than SVR and SEISMIC, which means that the L1 regularization comes into effect. As a probabilistic model, PreWhether gives a slightly poor detection result due to the assumption that all the features are independent. Though less effective than EPOC, SansNet outperforms all the other baselines in this classification task, since SansNet only employs one feature from cascades. However, it is plausible to note that SansNet gives stable k-coverage and Cost results in both Twitter and Weibo, which indicates that survival perspective models are suitable in this scenario.

Table 3. Result of burst detection on Twitter

		LR	SVR	PreWheter	SEISMIC	SansNet	EPOC
Twitter	**k-coverage**	0.7781	0.5969	0.7490	0.5188	0.8275	0.8471
	Cost	0.1032	0.0998	0.0956	0.1677	0.0776	0.0701
Weibo	**k-coverage**	0.6805	0.4918	0.6512	0.4589	0.7720	0.7784
	Cost	0.0951	0.1229	0.1271	0.1581	0.0961	0.0881

Change of Observation Periods. To explore the connection between observing period and the performance of methods, we conduct experiments on Twitter with six time periods from 0.5 to 3 hours and organize the results in Figure 5. Intuitively, the performances of EPOC and five baselines improve gradually as the observing period increases. We can clearly see that EPOC performs the best with a pretty high k-coverage at about 87% and a pretty low cost at around 0.068. Besides, it is worth noticing that SEISMIC is far behind other approaches no matter in k-coverage or in Cost, which suggests that time series model depends on a relatively longer observing period, and can not do a good job the burst detection task at early stage.

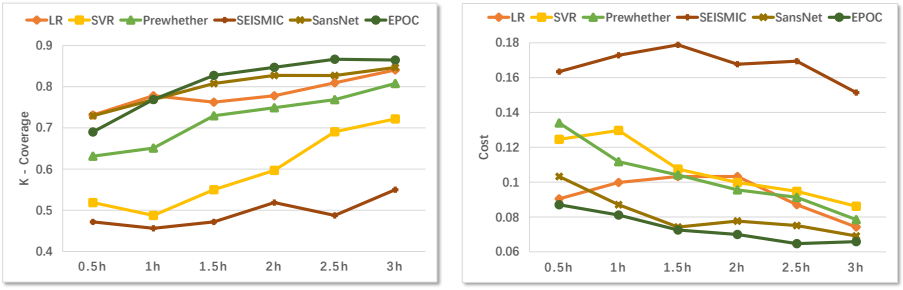


Fig. 5. k-Coverage and Cost under Different Observing Periods on Twitter

Time Ahead (similar to *EPA* from [8]): further, we try to figure out how early we can detect the outbreak cascades with EPOC. As [13] states, it is a pathological task to estimate the final size of a cascade if only given a short initial portion, since almost all cascades are small. Besides, comparing with getting the final size of a cascade, it is more meaningful and practical to detect how early a cascade will break out. Therefore, in this experiment of Twitter and Weibo,

we only probe into the early pattern of outbreak cascades, and mainly focus on *absolute time ahead*, which is the interval between the predicted burst time $t_{predict}$ and the actual burst time t_{actual}. Specifically during the experiments, if $t_{actual} \geq t_{predict}$, we record as $t_{actual} - t_{predict}$, and otherwise, 0. Also, we consider the *relative time ahead*, which is given by $\frac{t_{actual}-t_{predict}}{t_{actual}}$ or 0.

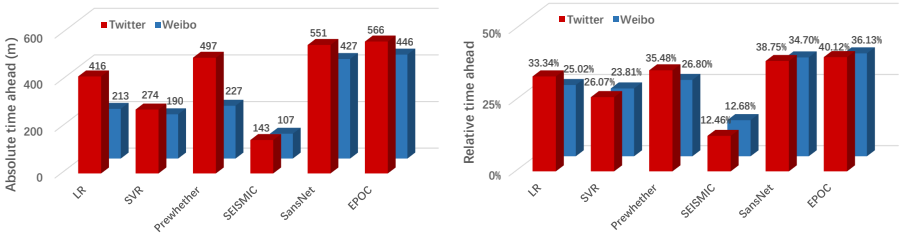


Fig. 6. Absolute and Relative Time Ahead on Twitter and Weibo

Performance Analysis. Figure 6 illustrates the corresponding experiment results on Twitter and Weibo. We conclude that all the methods have a similar rank in terms of absolute time ahead and relative time ahead. SansNet and our EPOC steadily keep a leading role in this regression task at about 38.75% and 40.12% respectively ahead of the actual burst time in Twitter. PreWhether and LR work mildly, and they can successfully predict the occurrence of burst, when the diffusion process of cascades only goes on about two thirds. Though SVR possesses much better performance than the poorest SEISMIC, it falls behind comparing with other baselines, which suggests that the notion of support vector may not be applicable in this problem.

5 Related Work

In recent years, social networks have successfully attracted researchers' attention, and plenty of achievements have been made in the past few decades, especially when it comes to the study of information cascades, including the prediction of cascade size, how the cascade grows and disseminates, etc.

5.1 Information Cascade and Social Networks

The study of information cascades has been going for a long time, and it is of great use in many applications, such as meme tracking [2], stock bubble diagnosis [3], and sales prediction [4]. The literature concerning cascade in social networks can be divided into three categories. The first category lays on user level prediction. One of the pioneers is Iwata et al. [18], they propose a Bayesian inference model with stochastic EM algorithm, trying to discover the latent influence among online users. [19] also utilizes user-related features to help social event detection. Additionally, some other researchers also analyze the topology, since structural feature is said to be one of the predictors of cascade size [13]. PageRank of retweeting graph is taken into consideration [20], while [21] utilizes

the number of directed followers as one of the important infectors. Another significant category is temporal features. Many experimental results, such as [10][9], reveal that temporal features are the most effective type of indicators. To depict the connection between early cascade and its final state, both [5] and [12] propose Bayesian networks with temporal information. Other temporal information, like mean time and maximum time interval, has also been considered [9].

5.2 Outbreak Detection and Modeling

Burst or outbreak, defined as "a brief period of intensive activity followed by long period of nothingness" [6], is a common phenomenon during the diffusion of social content, which is worthy of studying and may bring benefits to modern society. Existing works probing into cascades mainly focus on prediction of its future popularity [5][12][20] or final aggregate size [10][13]. However, how to detect the burst pattern of large cascade in early stage remains an intriguing problem. Recently, based on the transformation of time window, Wang et al. [6] proposes a classification model to predict the burst time of cascade. Unfortunately, their approach acquires laborious feature extraction, and the traditional classifiers they used can hardly take the best use of the features. [17] implements a logistic model, which considers all the nodes as cascade sensors. Just as bad, when the number of nodes in networks turns to be billions, the implementation of this method will be particularly difficult.

In this work, adopting survival theory, we can exactly overcome these drawbacks from the perspective of cascade dynamics. Other researchers also employ survival models to understand the burst of cascades. SansNet is proposed in [8], predicting whether and when a cascade goes viral. This approach utilizes only the size of cascades as feature, making it weak to apply to multiply cases, since the features of an author [22] and the inherent network [13] are sometimes more important than features from cascade itself [22]. Another drawback of this approach is that the survival curve cannot totally reveal the status of cascades.

6 Conclusion and Perspectives

In social networks, detecting whether and when a cascade will outbreak is a non-trivial but beneficial task. In this paper, we novelly employ survival theory, proposing a survival model EPOC to detect the early pattern of outbreak cascades. We extract both dynamic and static features from cascades and utilize Gaussian distributions to characterize their survival probabilities, then accompanied with hazard rate, we successfully detect the burst pattern of cascades at very early stage. Extensive experiment shows that our EPOC outperforms five state-of-the-art methods in this practical task.

As future work, firstly we will mainly concentrate on how to choose a better standard baseline for hazard ceiling, and more experiment observation might be made. Then, we will consider more influential and relevant features or try another suitable survival theory based model. Finally, we hope that our work will pave ways to richer and deeper understanding of cascades.

Acknowledgements. This work is supproted by the Program of International

S&T Cooperation (2016YFE0100300), the China 973 project (2014CB340303), the National Natural Science Foundation of China (61472252, 61672353), the Shanghai Science and Technology Fund (17510740200), and CCF-Tencent Open Research Fund (RAGR20170114).

References

1. Adrien, F., Lada, A., Dean, E., Justin C.: Rumor Cascades. In ICWSM (2014)
2. Bai, j., Li, L., Lu, L., Yang, Y., Zeng, D.: Real-time prediction of meme burst. In IEEE ISI (2017)
3. Jiang, Z., Zhou, W., Didier, S., Ryan, W., Ken, B., Peter, C.: Bubble Diagnosis and Prediction of the 2005–2007 and 2008–2009 Chinese Stock Market Bubbles. Journal of Economic Behavior & Organization (2010)
4. Daniel, G., Ramanathan, V. Ravi, K., Jasmine, N., Andrew, T.: The Predictive Power of Online Chatter. In SIGKDD (2005)
5. Ma, X., Gao, X., Chen, G.: BEEP: A Bayesian Perspective Early Stage Event Prediction Model for Online Social Networks. In ICDM (2017)
6. Wang, S., Yan, Z., Hu, X., Philip, S., Li, Z.: Burst Time Prediction in Cascades. In AAAI (2015)
7. Matsubara, Y., Sakurai, Y., Prakash, B., Li, L., Faloutsos C.: Rise and Fall Patterns of Information Diffusion: Model and Implications. In SIGKDD (2012)
8. Subbian, K., Prakash, B., Adamic, L.: Detecting Large Reshare Cascades in Social Networks. In WWW (2017)
9. Gao, S., Ma, J., Chen, Z.: Effective and Effortless Features for Popularity Prediction in Microblogging Network. In WWW (2014)
10. Zhao, Q., Erdogdu, M., He, H., Rajaraman, A., Leskovec, J.: SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In SIGKDD (2015)
11. Gao, S., Ma, J., Chen, Z.: Modeling and Predicting Retweeting Dynamics on Microblogging Platforms. In WSDM (2015)
12. Liu, W., Deng, Z, Gong, X., Jiang, F., Tsang, I.: Effectively Predicting Whether and When a Topic Will Become Prevalent in a Social Network. In AAAI (2015)
13. Cheng, J., Adamic, L., Dow, P., Kleinberg, J., Leskovec, J.: Can Cascades Be Predicted?. In WWW (2014)
14. Cox, R.: Regression Models and Life-Tables. Springer. Breakthroughs in Statistics (1992)
15. Aalen, O., Borgan, O., Gjessing, H.: Survival and Event History Analysis. Springer (2008)
16. Anderson, J., R., Bernstein, L., Pike, M., C.: Approximate Confidence Intervals for Probabilities of Survival and Quantiles in Life-Table Analysis. International Biometric Society. JSTOR. Vol. 38, No. 2 (1982)
17. Cui, P., Jin, S., Yu, L., Wang, F., Zhu, W., Yang, S.: Cascading Outbreak Prediction in Networks: a Data-Driven Approach. In SIGKDD (2013)
18. Iwata, T., Shah, A., Ghahramani, Z.: Discovering Latent Influence in Online Social Activities via Shared Cascade Poisson Processes. In SIGKDD (2013)
19. Mansour, E., Tekli, G., Arnould, P., Chbeir, R., Cardinale, Y.: F-SED: Feature-Centric Social Event Detection. In DEXA (2017)
20. Hong, L., Dan, O., Davison, B.: Predicting Popular Messages in Twitter. In WWW (2011)
21. Feng, Z., Li, Y., Jin, L., Feng, L.: A Cluster-Based Epidemic Model for Retweeting Trend Prediction on Micro-blog. In DEXA (2015)
22. Petrovic, S., Osborne, M., Lavrenko, V.: RT to Win! Predicting Message Propagation in Twitter. In ICWSM (2011)